



Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.

<http://siba-ese.unisalento.it/index.php/ejasa/index>

e-ISSN: 2070-5948

DOI: 10.1285/i20705948v7n1p58

**Identification of Multicollinearity and it's effect
in Model selection**

By Jayakumar and Sulthan

April 23, 2014

This work is copyrighted by Università del Salento, and is licensed under a Creative Commons Attribuzione - Non commerciale - Non opere derivate 3.0 Italia License.

For more information see:

<http://creativecommons.org/licenses/by-nc-nd/3.0/it/>

Identification of Multicollinearity and its effect in Model selection

D.S. Jayakumar * and A. Sulthan

¹*Jamal Institute of Management, Tiruchirappalli, India*

April 23, 2014

Multicollinearity is the problem experienced by statisticians while evaluating a sample regression model. This paper explored the relationship between the sample variance inflation factor (vif) and F-ratio, based on this we proposed an exact F-test for the identification of multicollinearity and it overcomes the traditional procedures of rule of thumb. The authors critically identified that the variance inflation factor not only inflates the variance of the estimated regression coefficient and it also inflates the residual error variance of a given fitted regression model in various level of inflation. Moreover, we also found a link between the problem of multicollinearity and its impact on the model selection decision. For this, the authors proposed multicollinearity corrected version of generalized information criteria which incorporates the effect of multicollinearity and help the statisticians to select a best model among the various competing models. This procedure numerically illustrated by fitting 12 different types of stepwise regression models based on 44 independent variables in a BSQ (Bank service Quality) study. Finally, the study result shows the transition in model selection after the correction of multicollinearity effect.

keywords: Multicollinearity, variance inflation factor, Error-variance, F-test, Generalized Information criteria, multicollinearity penalization.

1 Introduction and Related work

In the process of fitting regression model, when one independent variable is nearly combination of other independent variables, there will affect parameter estimates. This

*Corresponding authors: samjaya77@gmail.com

problem is called multicollinearity. Basically, multicollinearity is not a violation of the assumptions of regression but it may cause serious difficulties Neter et al (1989) (1) variances of parameter estimates may be unreasonably large, (2) parameter estimates may not be significant, (3) a parameter estimate may have a sign different from what is expected and so on Efron (2004). For solving or alleviating this problem in certain regression model, the usually best way is dropping redundant variables from this model directly, that is to try to avoid it by not including redundant variables in the regression model Bowerman et al (1993). But sometimes, it is hard to decide the redundant variables. Another alternative to deleting variables is to perform a principal component analysis Maddala (1977). With principal component regression, we create a set of artificial uncorrelated variables that can then be used in the regression model. Although principal component variables are dropped from the model, when the model is transformed back, it will cause other biases too Draper and Smith (1981) Srivastava (2002). The transformation of the independent variables and the methods applicable to overcome the multicollinearity problem discussed above purely depends on the exact identification of the problem. In this paper, the effect identification of the multicollinearity problem is discussed in a separate section and how it misleads the statisticians to select a regression model based on the information criteria are visualized in the next section.

2 Inflation of error variance

Consider an estimated sample regression model with a single dependent variable (y_i) with p regressors namely $x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi}$ is given as

$$y_i = \hat{\alpha} + \sum_{j=1}^p \hat{\beta}_j x_{ji} + \hat{e}_i \quad (1)$$

where $\hat{\alpha}$ is the estimated Intercept, $\hat{\beta}_j$ is the estimated beta co-efficients or partial regression co-efficients and \hat{e}_i is the estimated residual followed normal distribution $N(0, \sigma_e^2)$. From (1), the sample regression model should satisfy the assumptions of normality, homoscedasticity of the error variance and the serial independence property. Though the model satisfying all the assumptions, still it has to be evaluated. The authors more particularly focused on the multicollinearity and its effects leads the strong inter causal effect among the independent variables. For the past 5 decades, statisticians believe that the impact of this multicollinearity problem severely inflates the variance of the estimated regression co-efficients. This creates greater instability and inconsistency in the estimated co-efficients. Besides this, we identified a remarkable and astonishing problem due to the multicollinearity and it will be mathematically identified below. Consider the variance of the estimated regression co-efficient as

$$\widehat{\sigma_{\beta_j}^2} = \frac{s_e^2}{(n-1)s_{x_j}^2} \left(\frac{1}{1 - R_{x_j}^2} \right) \quad (2)$$

Where s_e^2 is the unbiased estimate of the error variance, $s_{x_j}^2$ is the variance of the x_j independent variable ($j=1,2,3 \dots p$) and $1/1 - R_{x_j}^2$ is technically called as variance inflation factor (vif). The term $1 - R_{x_j}^2$ is the unexplained variation in the x_j independent variable due to the same independent variables other than x_j . More specifically, statisticians named the term as Tolerance and inverse of the Tolerance is said to be the VIF. By using the fact $s_e^2 = (n/n - k)\widehat{\sigma_e^2}$ Rewrite (2) as

$$\widehat{\sigma_{\beta_j}^2} = \frac{n}{(n-1)(n-k)s_{x_j}^2} \left(\frac{\widehat{\sigma_e^2}}{1 - R_{x_j}^2} \right) \quad (3)$$

$$\widehat{\sigma_{\beta_j}^2} = \frac{n}{(n-1)(n-k)s_{x_j}^2} \widehat{\sigma_{INF(e_j)}^2} \quad (4)$$

From (3), the error variance ($\widehat{\sigma_e^2}$) of the given regression model plays a mediating role between the variance of estimated regression co-efficients ($\widehat{\sigma_{\beta_j}^2}$) and the VIF. Instead of analyzing the inflation of variance of estimated regression co-efficients ($\widehat{\sigma_{\beta_j}^2}$), the authors only focused on the inflated part of the error variance due to the impact of multicollinearity as from (4). From (4) $\widehat{\sigma_{INF(e_j)}^2}$ is the inflated error variance which is inflated by the $(VIF)_j$ is equal to

$$\widehat{\sigma_{INF(e_j)}^2} = \frac{\widehat{\sigma_e^2}}{1 - R_{x_j}^2} \quad (5)$$

$$\widehat{\sigma_{INF(e_j)}^2} = \frac{\sum_{i=1}^n (\widehat{e_i} / \sqrt{1 - R_{x_j}^2})^2}{n} \quad (6)$$

$$\widehat{\sigma_{INF(e_j)}^2} = \frac{\sum_{i=1}^n (\widehat{e_{INF(e_{ji})}})^2}{n} \quad (7)$$

From (5), the inflated error variance $\widehat{\sigma_{INF(e_j)}^2}$ which is always greater than or equal to the uninflated error variance $\widehat{\sigma_e^2}$ where $(\widehat{\sigma_{INF(e_j)}^2} \geq \widehat{\sigma_e^2})$. If $R_{x_j}^2$ is equal to 0, then both the variances are equal, there is no multicollinearity. Similarly, If the $R_{x_j}^2$ is equal to 1, then the error variance severely inflated and raise upto 8 and this shows the existence of severe multicollinearity. In the same manner, if $0 < R_{x_j}^2 < 1$, then there will be a chance of inflation in the error variance of the regression model. Likewise, from (6) and (7), the estimated errors $\widehat{e_i}$ are also inflated by the $\sqrt{(VIF)_j}$ and it is transformed as estimated inflated residuals $\widehat{e_{INF(e_{ji})}}$. If the estimated errors and error variance are inflated, then the forecasting performance of the model will decline and this leads to take inappropriate and illogic model selection decision. From the above discussion, the authors proved the problem of multicollinearity not only inflates the variance of estimated regression co-efficients but also inflates the error variance of the regression model. In order to find the

statistical equality between the inflated error variance $\widehat{\sigma_{INF(e_j)}^2}$ and the uninflated error variance $\widehat{\sigma_e^2}$, the authors proposed an F-test by finding the link between sample vif and F-ratio. The methodology of applying the test statistic is discussed in the next section.

3 Testing the Inflation of error variance

For the purpose of testing the statistical equality between sample $\widehat{\sigma_e^2}$ and $\widehat{\sigma_{INF(e_j)}^2}$, first, the authors derived the test statistic by re-writing (5) as the basis and it is given as

$$\frac{1}{1 - R_{x_j}^2} = \frac{\widehat{\sigma_{INF(e_j)}^2}}{\widehat{\sigma_e^2}} \quad (8)$$

$$(vif)_j = \frac{\widehat{\sigma_{INF(e_j)}^2}}{\widehat{\sigma_e^2}} \quad (9)$$

From (8) and (9), it has been modified as

$$\frac{R_{x_j}^2}{1 - R_{x_j}^2} = \frac{\widehat{\sigma_{INF(e_j)}^2}}{\widehat{\sigma_e^2}} - 1 \quad (10)$$

$$\frac{R_{x_j}^2}{1 - R_{x_j}^2} = (vif)_j - 1 \quad (11)$$

From (11), we using the fact $(sst)_j = (ssr)_j + (sse)_j$, $R_{x_j}^2 = (ssr)_j / (sst)_j$, $1 - R_{x_j}^2 = (sse)_j / (sst)_j$, rewrite (11) as

$$\frac{(ssr)_j}{(sse)_j} = (vif)_j - 1 \quad (12)$$

Where ssr , sse , sst refers to the sample sum of squares of regression, error and the total respectively. Based on (12), it can be rewritten as in terms of the sample mean squares of regression (s_r^2) and error (s_e^2) as

$$\frac{qs_{r_j}^2}{(n - q - 1)s_{e_j}^2} = ((vif)_j - 1) \quad (13)$$

From (13), multiply both sides by the population mean square ratios $\sigma_{e_j}^2 / \sigma_{r_j}^2$, we get

$$\frac{qs_{r_j}^2 / \sigma_{r_j}^2}{(n - q - 1)s_{e_j}^2 / \sigma_{e_j}^2} = \left(\frac{\sigma_{e_j}^2}{\sigma_{r_j}^2}\right)((vif)_j - 1) \quad (14)$$

From (14), the ratios $qs_{r_j}^2 / \sigma_{r_j}^2$ and $(n - q - 1)s_{e_j}^2 / \sigma_{e_j}^2$ are followed chi-square distribution with q and $n - q - 1$ degrees of freedom (where q is the no. of independent variables in the auxiliary regression model $x_j = \alpha_{0j} + \sum_{k=1}^q \alpha_{jk}x_k + e_j, j \neq k$) and they are independent.

The independency of the ratios are based on least square property if $x_{ji} = \widehat{x}_{ji} + e_{ji}$, then \widehat{x}_{ji} and e_{ji} are independent and the respective sum of squares are equal to $(sst)_j = (ssr)_j + (sse)_j$. without loss of generality, (14) can be written as in terms of the F-ratio and it is defined as the ratio between the two independent chi-square variates divided by the respective degrees of freedom and is given as

$$\frac{\chi_{r_j}^2/q}{\chi_{e_j}^2/(n-q-1)} = \frac{(n-q-1)\sigma_{e_j}^2}{q\sigma_{r_j}^2}((vif)_j - 1) \quad (15)$$

$$F_{j(q,n-q-1)} = \left(\frac{n-q-1}{q}\right)((vif)_j - 1)\left(\frac{\sigma_{e_j}^2}{\sigma_{r_j}^2}\right) \sim F_{(q,n-q-1)} \quad (16)$$

From (16), the population variance ratios $\sigma_{e_j}^2/\sigma_{r_j}^2$ can be written in terms of the population VIF by combining the ratios $\sigma_{e_j}^2/\sigma_{r_j}^2$ and (12) can be given as

$$\frac{\sigma_{r_j}^2}{\sigma_{e_j}^2} = (VIF)_j - 1 \quad (17)$$

$$\frac{\sigma_{e_j}^2}{\sigma_{r_j}^2} = \frac{1}{(VIF)_j - 1} \quad (18)$$

Then, substitute (18) in (16), we get the final version of the F-ratio of the j th independent variable using in the regression model as

$$F_{j(q,n-q-1)} = \frac{(n-q-1)}{q} \left(\frac{(vif)_j - 1}{(VIF)_j - 1} \right) \sim F_{(q,n-q-1)} \quad (19)$$

From (19), the relationship between the sample vif and the F-ratio can be derived as

$$(vif)_j = 1 + ((VIF)_j - 1) \frac{q}{(n-q-1)} F_j \quad (20)$$

This relationship equation is the significant part of this research paper and the authors believe, it can be used to test the exactness of multi-collinearity and its effect on the error variance of regression model. The sampling distribution of vif exists, if the population $VIF > 1$ and if $VIF=1$, then the sample $vif=1$. under the null hypothesis, if the population $VIF=2$, then it shows to check the cent percent inflation in the population error variances of a model and the test statistic from (19) is given as

$$F_{j(q,n-q-1)} = \frac{(n-q-1)}{q} ((vif)_j - 1) \sim F_{(q,n-q-1)} \quad (21)$$

Where from (21), vif is the sample variance factor of j th independent variable used in a regression model, q is the no. of independent variables in the auxillary regression model and n is the sample size. Pragmatically, the authors recommends for checking the variance inflation in different levels gives more exact results. If $VIF=2$, then from (9), the population inflated error variance is equal to twice of the population error

variance, that is $\sigma_{INF(e_j)}^2 = 2\sigma_e^2$. This shows, due to the multicollinearity effect, the error variance of regression model is inflated 100% percent. The authors believe, the exact test of multicollinearity and its effect can be determined by fixing the value of VIF as $1 < (VIF)_j \leq 2$. If $VIF=1.05, 1.1, 1.5, 2$ helps the statisticians to scrutinize the inflation in error variance due to the multicollinearity effect at 5%, 10%, 50% and 100% inflation level.

4 Modified Generalized Information Criterion and Incorporation of Multicollinearity effect in Model selection

In the previous sections, the authors discussed, how the vif inflates the estimated error variance of a model as well as they also highlighted the application of the F -test to diagnose the exact impact of the multicollinearity with different inflation levels due to the independent variables in a regression model. This section deals with the model selection aspect, and it leads to modify the existing model selection criteria, which the selection criteria should carry the multicollinearity effect and the authors tactically penalize the models, whenever the selected model is having multicollinearity problem. For a multiple regression model, the generalized information criterion is given as

$$GIC = n \log(\widehat{\sigma_e^2}) + f(n, k) \quad (22)$$

where $\widehat{\sigma_e^2}$ the estimated error variance of the regression is model and $f(n, k)$ is the generalized penalty function which includes the sample size 'n' and no. of parameters 'k' estimated in a regression model. In order to incorporate the effect of multicollinearity effect in GIC, the authors highlighted some procedure as follows.

Step: 1 Calculate the residuals \widehat{e}_i for i th observation from a fitted sample regression model, where $e_i \sim N(0, \sigma_e^2)$.

Step: 2 Inflate the residuals calculated from step 1 by using the square root of the Average variance inflation factor (\sqrt{AVIF}), where $AVIF = \frac{1}{m} \sum_{j=1}^m (1/1 - R_j^2) = \frac{1}{m} \sum_{j=1}^m (vif)_j$ and m is the no. of significant sample vif based on the F-test of significance from (19).

Step: 3 Calculate the inflated residuals $INF(\widehat{e}_i)$ for i th observation from step 2, where $INF(\widehat{e}_i) \sim N(0, \sigma_{INF(e)}^2)$.

Step: 4 Estimate the inflated error variance from $INF(\widehat{e}_i) = \widehat{e}_i \sqrt{AVIF}$ and we get $\widehat{\sigma_{INF(e)}^2} = \widehat{\sigma_e^2}(AVIF)$.

Step: 5 From step 4, rewrite $\widehat{\sigma_{INF(e)}^2} = \widehat{\sigma_e^2}(AVIF)$ as $\widehat{\sigma_e^2} = \widehat{\sigma_{INF(e)}^2}/AVIF$ and substitute in (18), we get the Modified Generalized information criterion ($MGIC$)

Based on the above discussed procedure, the multicollinearity effect is incorporated into the existing information criterion (9) is given as

$$\begin{aligned}
GIC &= n \log(\widehat{\frac{\sigma_{INF(e)}^2}{AVIF}}) + f(n, k) \\
GIC &= n \log(\widehat{\sigma_{INF(e)}^2}) - n \log(AVIF) + f(n, k) \\
GIC + n \log(AVIF) &= n \log(\widehat{\sigma_{INF(e)}^2}) + f(n, k) \\
MGIC &= n \log(\widehat{\sigma_{INF(e)}^2}) + f(n, k)
\end{aligned} \tag{23}$$

where $MGIC$ is the Modified Generalized information criterion, $\widehat{\sigma_{INF(e)}^2}$ is the inflated estimate of population error variance and $f(n, k)$ is the penalty function. Moreover, this $MGIC$ is more meaningful when compared it with the traditional information criteria because the error variance was inflated and it will be penalized due to the multicollinearity effect existing in a regression model. The authors also derived the Akaike information criterion (AIC), Schwartz Bayesian criterion (SBIC), Hannan-Quinn information criterion, Akaike information criterion corrected for small sample (AICc) from the Modified Generalized information criterion ($MGIC$) by assigning the proper values for the penalty function which leads us to get 4 different modified version of the AIC, SBIC, HQIC and AICc. These criteria can also be used for the purpose of selecting a best model whenever the problem of multicollinearity exists among the set of competing models. The following table shows the different versions of information criteria after modification.

Table 1: Modified versions of Generalized Information Criteria

Criteria	Penalty function	Before multicoll. correction	After multicollinearity
GIC	$f(n, k)$	$n \log(\widehat{\sigma_e^2}) + f(n, k)$	$n \log(\widehat{\sigma_{INF(e)}^2}) + f(n, k)$
AIC	$2k$	$n \log(\widehat{\sigma_e^2}) + 2k$	$n \log(\widehat{\sigma_{INF(e)}^2}) + 2k$
SBIC	$k \log n$	$n \log(\widehat{\sigma_e^2}) + k \log n$	$n \log(\widehat{\sigma_{INF(e)}^2}) + k \log n$
HQIC	$2k \log(\log n)$	$n \log(\widehat{\sigma_e^2}) + 2k \log(\log n)$	$n \log(\widehat{\sigma_{INF(e)}^2}) + 2k \log(\log n)$
AICc	$2nk/n - k - 1$	$n \log(\widehat{\sigma_e^2}) + (2nk/n - k - 1)$	$n \log(\widehat{\sigma_{INF(e)}^2}) + (2nk/n - k - 1)$

k- no. of parameters (includes intercept + error variance), n- sample size, $\widehat{\sigma_e^2}$ - Sample estimate of error variance, $\widehat{\sigma_{INF(e)}^2}$ -Inflated sample estimate of error variance

5 Results and Discussion

In this section, we will investigate the effect of multicollinearity in the model selection decision based on the survey data collected from BSQ (Bank service quality) study. The

data comprised of 45 different service quality attributes about the bank and the data was collected from 102 account holders. A well-structured questionnaire was prepared and distributed to 125 customers and the questions were anchored at 5 point likert scale from 1 to 5 after the data collection was over only 102 completed questionnaires were used for analysis. The aim of this article is to identify the exactness of multicollinearity and how it inflates the error variance of a fitted regression model. Moreover the authors also discussed how the multicollinearity distorts the model selection decision as well as the transition occurred in model selection. The following table shows the results extracted from the analysis by using IBM SPSS version 21. At first the authors used stepwise multiple regression analysis by utilizing 44 independent variables and a dependent variable. The results of the stepwise regression analysis, test of multicollinearity and the transitions in model selection are visualized in the following tables.

Table 2: F -test for Multicollinearity identification and inflation of error variance

Independent variables	Model 2					Model 3										
	Tol.	vif	$\widehat{\sigma_{INF(e)}^2}$	F-ratios Variance Inflation level				Tol.	Vif	$\widehat{\sigma_{INF(e)}^2}$	F-ratiob Variance Inflation level					
X38	0.993	1.007	0.191	5%	10%	50%	100%		.992	1.008	0.178	5%	10%	Limit	50%	100%
X2	0.993	1.007	0.191	14.00**	7.00**	1.40	0.70		.982	1.018	0.180	17.82**	8.91**		1.78	0.89
X31	-	-	-	-	-	-	-		.989	1.011	0.179	10.89**	5.44**		1.09	0.54
X19	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X44	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X5	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X11	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X13	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X39	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X8	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
X23	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
error variance	$\widehat{\sigma_e^2}=0.18981$								$\widehat{\sigma_e^2}=0.17679$							

i d.f.(9,92), ** p-value<0.01, * p-value<0.05, j d.f.(10,91), ** p-value<0.01, * p-value<0.05

Table 3: F-test for Multicollinearity identification and inflation of error variance

Independent variables		Model 4				Model 5									
		Tol.	vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratioc Variance Inflation level			Tol.	vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratioc Variance Inflation level				
					5%	10%	50%	100%			5%	10%	50%	100%	
X38		.978	1.023	0.171	15.03**	7.51**	1.50	.75	.864	1.158	0.190	76.63**	38.31**	7.66**	3.83**
X2		.967	1.035	0.173	22.87**	11.43**	2.29	1.14	.964	1.037	0.170	17.94**	8.97**	1.79	0.90
X31		.973	1.028	0.172	18.29**	9.15**	1.83	0.91	.951	1.051	0.173	24.73**	12.37**	2.47*	1.24
X19		.949	1.054	0.176	35.28	17.64**	3.53*	1.76	.947	1.056	0.173	27.16**	13.58**	2.72*	1.36
X44		-	-	-	-	-	-	-	.861	1.161	0.191	78.08**	39.04**	7.81**	3.90**
X5		-	-	-	-	-	-	-	-	-	-	-	-	-	-
X11		-	-	-	-	-	-	-	-	-	-	-	-	-	-
X13		-	-	-	-	-	-	-	-	-	-	-	-	-	-
X39		-	-	-	-	-	-	-	-	-	-	-	-	-	-
X8		-	-	-	-	-	-	-	-	-	-	-	-	-	-
X23		-	-	-	-	-	-	-	-	-	-	-	-	-	-
error variance		$\widehat{\sigma^2_e}=0.17679$				$\widehat{\sigma^2_e}=0.16422$									

Table 4: F -test for Multicollinearity identification and inflation of error variance

Independent variables		Model 4				Model 5								
	Tol.	vif	$\widehat{\sigma_{INF(\epsilon)}^2}$	F-ratioe Variance Inflation level				Tol.	vif	$\widehat{\sigma_{INF(\epsilon)}^2}$	F-ratioe Variance Inflation level			
				5%	10%	50%	100%				5%	10%	50%	100%
X38	.978	1.023	0.171	15.03**	7.51**	1.50	.75	.864	1.158	0.190	76.63**	38.31**	7.66**	3.83**
X2	.967	1.035	0.173	22.87**	11.43**	2.29	1.14	.964	1.037	0.170	17.94**	8.97**	1.79	0.90
X31	.973	1.028	0.172	18.29**	9.15**	1.83	0.91	.951	1.051	0.173	24.73**	12.37**	2.47*	1.24
X19	.949	1.054	0.176	35.28	17.64**	3.53*	1.76	.947	1.056	0.173	27.16**	13.58**	2.72*	1.36
X44	-	-	-	-	-	-	-	.861	1.161	0.191	78.08**	39.04**	7.81**	3.90**
X5	-	-	-	-	-	-	-	-	-	-	-	-	-	-
X11	-	-	-	-	-	-	-	-	-	-	-	-	-	-
X13	-	-	-	-	-	-	-	-	-	-	-	-	-	-
X39	-	-	-	-	-	-	-	-	-	-	-	-	-	-
X8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
X23	-	-	-	-	-	-	-	-	-	-	-	-	-	-
error variance	$\widehat{\sigma_e^2}=0.17679$							$\widehat{\sigma_e^2}=0.16422$						

C d.f.(3,98), ** p-value<0.01, * p-value<0.05, d d.f.(4,97), ** p-value<0.01, * p-value<0.05

Table 5: F -test for Multicollinearity identification and inflation of error variance

Independent variables		Model 6			Model 7		
		Tol.	vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratio Variance Inflation level		F-ratio Variance Inflation level
					5%	10%	5%
X38		.860	1.163	0.182	62.59**	31.30**	52.57**
X2		.950	1.053	0.165	20.35**	10.18**	54.78**
X31		.951	1.052	0.165	19.97**	9.98**	25.65**
X19		.931	1.074	0.168	28.42**	14.21**	37.05**
X44		.815	1.226	0.192	86.78**	43.39**	78.22**
X5		.921	1.086	0.170	33.02**	16.51**	34.83**
X11		-	-	-	-	-	89.30**
X13		-	-	-	-	-	-
X39		-	-	-	-	-	-
X8		-	-	-	-	-	-
X23		-	-	-	-	-	-
error variance		$\widehat{\sigma^2}=0.15650$			$\widehat{\sigma^2}=0.14721$		

e d.f(5,96), ** p-value<0.01, * p-value<0.05, ** p-value<0.01, * p-value<0.05

Table 6: F -test for Multicollinearity identification and inflation of error variance

Independent variables		Model 8				Model 9											
		Tol.	vif	$\widehat{\sigma^2_{INF(\epsilon)}}$	F-ratio ^g				Tol.	vif	$\widehat{\sigma^2_{INF(\epsilon)}}$	F-ratio ^h					
					Variance	inflation level						Variance	inflation level				
					5%	10%	50%	100%				5%	10%	50%	100%		
X38		.848	1.180	0.166	48.34**	24.17**	4.83**	2.42*	.826	1.210	0.161	48.82**	24.41**	4.88**	2.44*		
X2		.852	1.174	0.166	46.73**	23.37**	4.67**	2.34*	.808	1.238	0.165	55.33**	27.67**	5.53**	2.77**		
X31		.906	1.103	0.156	27.66**	13.83**	2.77*	1.38	.803	1.245	0.166	56.96**	28.48**	5.70**	2.85**		
X19		.854	1.171	0.165	45.93**	22.96**	4.59**	2.30*	.836	1.197	0.159	45.80**	22.90**	4.58**	2.29*		
X44		.797	1.255	0.177	68.49**	34.24**	6.85**	3.42**	.797	1.255	0.167	59.29**	29.64**	5.93**	2.96**		
X5		.895	1.118	0.158	31.69**	15.85**	3.17**	1.58	.872	1.147	0.153	34.18**	17.09**	3.42**	1.71		
X11		.757	1.321	0.186	86.21**	43.11**	8.62**	4.31**	.755	1.325	0.176	75.56**	37.78**	7.56**	3.78**		
X13		.857	1.167	0.165	44.85**	22.43**	4.49**	2.24*	.855	1.169	0.156	39.29**	19.65**	3.93**	1.96		
X39		-	-	-	-	-	-	-	.796	1.257	0.167	59.75**	29.88**	5.98**	2.99**		
X8		-	-	-	-	-	-	-	-	-	-	-	-	-	-		
X23		-	-	-	-	-	-	-	-	-	-	-	-	-	-		
error variance		$\widehat{\sigma^2_{\epsilon}}=0.14099$								$\widehat{\sigma^2_{\epsilon}}=0.13317$							

g d.f(7,94), ** p-value<0.01, * p-value<0.05, h d.f(8,93), ** p-value<0.01, * p-value<0.05

Table 7: *F*-test for Multicollinearity identification and inflation of error variance

Journal of Applied Statistical

Independent variables		Model 10				Model 11											
		Tol.	Vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratioi Variance inflation level				Tol.	vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratioj Variance inflation level					
					5%	10%	50%	100%				5%	10%	50%	100%		
X38		.811	1.233	0.155	47.64**	23.82**	4.76**	2.38*	.799	1.251	0.154	45.68**	22.84**	4.57**	2.28*		
X2		.806	1.241	0.156	49.27**	24.64**	4.93**	2.46*	.748	1.337	0.165	61.33**	30.67**	6.13**	3.07**		
X31		.802	1.248	0.157	50.70**	25.35**	5.07**	2.54*	.673	1.485	0.183	88.27**	44.13**	8.83**	4.41**		
X19		.835	1.197	0.150	40.28**	20.14**	4.03**	2.01*	.800	1.250	0.154	45.50**	22.75**	4.55**	2.28*		
X44		.781	1.280	0.161	57.24**	28.62**	5.72**	2.86**	.760	1.315	0.162	57.33**	28.66**	5.73**	2.87**		
X5		.864	1.158	0.146	32.30**	16.15**	3.23**	1.62	.864	1.158	0.143	28.76**	14.38**	2.88**	1.44		
X11		.665	1.503	0.189	102.84**	51.42**	10.28**	5.14**	.665	1.504	0.186	91.73**	45.86**	9.17**	4.59**		
X13		.855	1.169	0.147	34.55**	17.28**	3.46**	1.73	.844	1.184	0.146	33.49**	16.74**	3.35**	1.67		
X39		.726	1.378	0.173	77.28**	38.64**	7.73**	3.86**	.724	1.381	0.170	69.34**	34.67**	6.93**	3.47**		
X8		.753	1.327	0.167	66.85**	33.43**	6.69**	3.34**	.704	1.420	0.175	76.44**	38.22**	7.64**	3.82**		
X23		-	-	-	-	-	-	-	.651	1.536	0.189	97.55**	48.78**	9.76**	4.88**		
error variance		$\widehat{\sigma^2_e}=0.12569$								$\widehat{\sigma^2_e}=0.123078$							

i d.f.(9,92), ** p-value<0.01, * p-value<0.05, j d.f.(10,91), ** p-value<0.01, * p-value<0.05

Table 8: *F*-test for Multicollinearity identification and inflation of error variance

Independent variables		Model 12					
	Tol.	vif	$\widehat{\sigma^2_{INF(e)}}$	F-ratio k			
				Variance inflation level			
				5%	10%	50%	100%
X38	.811	1.233	0.157	47.64**	23.82**	4.76**	2.38*
X2	.748	1.336	0.170	68.69**	34.35**	6.87**	3.43**
X31	-	-	-	-	-	-	-
X19	.806	1.241	0.158	49.27**	24.64**	4.93**	2.46*
X44	.789	1.268	0.161	54.79**	27.40**	5.48**	2.74**
X5	.866	1.155	0.147	31.69**	15.84**	3.17**	1.58
X11	.681	1.469	0.187	95.88**	47.94**	9.59**	4.79**
X13	.851	1.175	0.149	35.78**	17.89**	3.58**	1.79
X39	.790	1.266	0.161	54.38**	27.19**	5.44**	2.72**
X8	.707	1.415	0.180	84.84**	42.42**	8.48**	4.24**
X23	.775	1.290	0.164	59.29**	29.64**	5.93**	2.96**
error variance		$\widehat{\sigma^2_e}=0.127046$					

kd.f(9,92), ** p-value<0.01, * p-value<0.05

Table 9: Transition in model selection at 5% Inflation level of error variance

Model	k	R2	Significant	$\widehat{\sigma^2_{INF(e)}}$	AIC		SBIC		HQIC		AICc	
					BMC	AMC	BMC	AMC	BMC	AMC	BMC	AMC
2	4	.331	1.007	0.19114	-161.50	-160.78	-151.00	-150.28	-157.24	-156.53	-161.08	-160.37
3	5	.377	1.0123	0.17896	-166.74	-165.50	-153.62	-152.38	-161.43	-160.19	-166.12	-164.88
4	6	.410	1.0286	0.18184	-164.74	-161.87	-149.00	-146.12	-158.37	-155.49	-163.86	-160.99
5	7	.441	1.0926	0.17943	-170.27	-161.23	-151.89	-142.86	-162.83	-153.79	-169.08	-160.04
6	8	.489	1.109	0.17356	-173.18	-162.63	-152.18	-141.63	-164.68	-154.12	-171.63	-161.08
7	9	.525	1.168	0.17194	-177.42	-161.58	-153.80	-137.96	-167.85	-152.02	-175.46	-159.63
8	10	.565	1.1861	0.16723	-179.82	-162.42	-153.58	-136.17	-169.20	-151.79	-177.41	-160.00
9	11	.598	1.227	0.16340	-183.65	-162.78	-154.77	-133.90	-171.95	-151.09	-180.71	-159.85
10	12	.615	1.2734	0.16005	-187.54	-162.89	-156.04	-131.39	-174.79	-150.14	-184.04	-159.39
11	13	.634	1.3473	0.16582	-187.68	-157.28	-153.56	-123.15	-173.87	-143.46	-183.55	-153.14
12	12	.630	1.2848	0.16323	-186.45	-160.88	-154.95	-129.39	-173.69	-148.13	-182.94	-157.38

BMC- Before multicollinearity correction, k-no.of parameters,avif- average variance inflation factor, AMC- After multicollinearity correction

Table 10: Transition in model selection at 10% Inflation level of error variance

Model	k	R ²	Significantavif	$\sigma_{INF(c)}^2$	BMC	AMC	SBIC	AMC	BMC	AMC	HQIC	BMC	AMC	BMC	AMC	AICc
2	4	.331	1.007	.19114	-161.50	-160.78	-151.00	-150.28	-157.24	-156.53	-161.08	-160.37				
3	5	.377	1.0123	.17896	-166.74	-165.50	-153.62	-152.38	-161.43	-160.19	-166.12	-164.88				
4	6	.410	1.035	.18298	-164.74	-161.23	-149.00	-145.48	-158.37	-154.86	-163.86	-160.35				
5	7	.441	1.0926	.17943	-170.27	-161.23	-151.89	-142.86	-162.83	-153.79	-169.08	-160.04				
6	8	.489	1.109	.17356	-173.18	-162.63	-152.18	-141.63	-164.68	-154.12	-171.63	-161.08				
7	9	.525	1.168	.17194	-177.42	-161.58	-153.80	-137.96	-167.85	-152.02	-175.46	-159.63				
8	10	.565	1.1861	.16723	-179.82	-162.42	-153.58	-136.17	-169.20	-151.79	-177.41	-160.00				
9	11	.598	1.227	.16340	-183.65	-162.78	-154.77	-133.90	-171.95	-151.09	-180.71	-159.85				
10	12	.615	1.2734	.16005	-187.54	-162.89	-156.04	-131.39	-174.79	-150.14	-184.04	-159.39				
11	13	.634	1.3473	.16582	-187.68	-157.28	-153.56	-123.15	-173.87	-143.46	-183.55	-153.14				
12	12	.630	1.2848	.16323	-186.45	-160.88	-154.95	-129.39	-173.69	-148.13	-182.94	-157.38				

BMC- Before multicollinearity correction, k-no.of parameters avif- average variance inflation factor, AMC- After multicollinearity correction

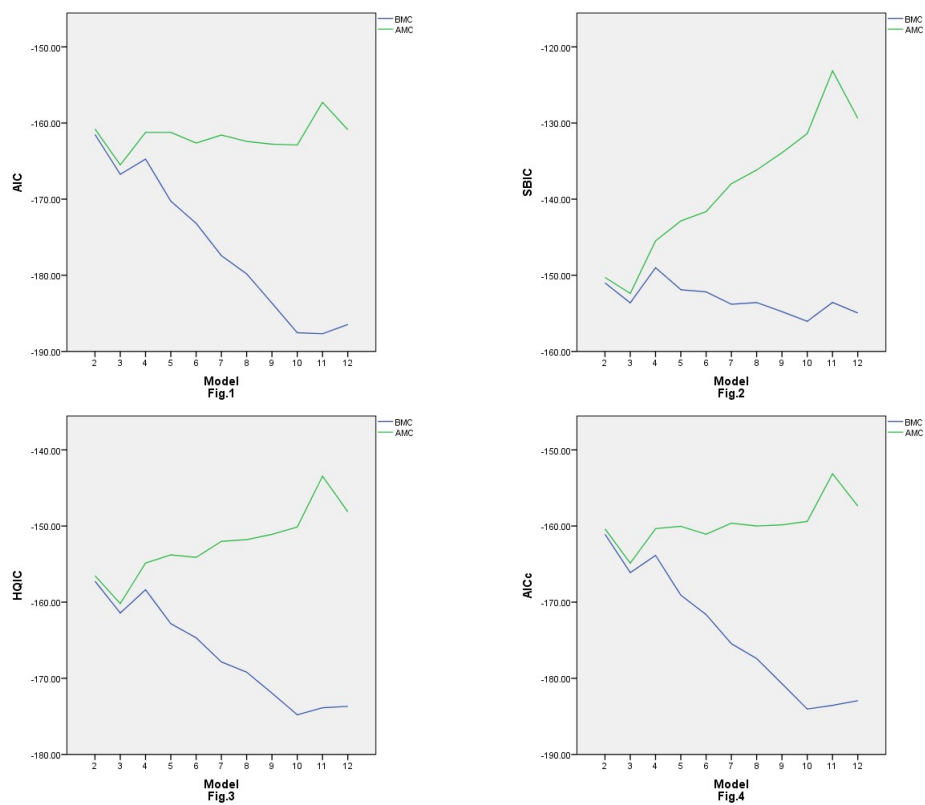


Figure 1: Line plot shows the values of Modified Versions of Information Criteria for different Models at 5% and 10% inflation in error variance

Table 11: Transition in model selection at 50% Inflation level of error variance

Model	k	R2	Significant	vif	$\sigma^2_{N F(e)}$	SBIC		HQIC		AICc	
						BMC	AMC	BMC	AMC	BMC	AMC
2	4	.331	-		.18981	-161.50	-161.50				
3	5	.377	-		.17679	-166.74	-166.74				
4	6	.410	1.054		.18633	-164.74	-159.38				
5	7	.441	1.1065		.18170	-170.27	-159.95				
6	8	.489	1.1372		.17798	-173.18	-160.06				
7	9	.525	1.168		.17194	-177.42	-161.58				
8	10	.565	1.1861		.16722	-179.82	-162.42				
9	11	.598	1.227		.16340	-183.65	-162.78				
10	12	.615	1.2734		.16005	-187.54	-162.89				
11	13	.634	1.3473		.16582	-187.68	-157.28				
12	12	.630	1.2848		.16323	-186.45	-160.88				

BMC- Before multicollinearity correction, k-no. of parameters, avf- average variance inflation factor, AMC- After multicollinearity correction

Table 12: Transition in model selection at 100% Inflation level of error variance

Model	k	R2	Significant	$\widehat{\sigma^2_{INF(e)}}$	AIC		SBIC		HQIC		AICc	
					BMC	AMC	BMC	AMC	BMC	AMC	BMC	AMC
2	4	.331	-	.18981	-161.50	-161.50	-151.00	-151.00	-157.24	-157.24	-161.08	-161.08
3	5	.377	-	.17679	-166.74	-166.74	-153.62	-153.62	-161.43	-161.43	-166.12	-166.12
4	6	.410	-	.17679	-164.74	-164.74	-149.00	-149.00	-158.37	-158.37	-163.86	-163.86
5	7	.441	1.1595	.19041	-170.27	-155.17	-151.89	-136.80	-162.83	-147.73	-169.08	-153.98
6	8	.489	1.1945	.18694	-173.18	-155.05	-152.18	-134.05	-164.68	-146.55	-171.63	-153.50
7	9	.525	1.217	.17915	-177.42	-157.39	-153.80	-133.77	-167.85	-147.83	-175.46	-155.44
8	10	.565	1.2113	.17078	-179.82	-160.27	-153.58	-134.02	-169.20	-149.64	-177.41	-157.86
9	11	.598	1.2467	.16602	-183.65	-161.16	-154.77	-132.28	-171.95	-149.46	-180.71	-158.22
10	12	.615	1.3008	.16349	-187.54	-160.72	-156.04	-129.22	-174.79	-147.97	-184.04	-157.22
11	13	.634	1.3865	.17065	-187.68	-154.35	-153.56	-120.23	-173.87	-140.53	-183.55	-150.21
12	12	.630	1.3147	.16703	-186.45	-158.54	-154.95	-127.04	-173.69	-145.78	-182.94	-155.03

BMC- Before multicollinearity correction, k-no.of parameters,avif- average variance inflation factor, AMC- After multicollinearity correction

Table 13: Information Loss based on Variance Inflation level

Model	p	AIC				SBIC				HQIC				AICc			
		Variance Inflation level				Variance Inflation level				Variance Inflation level				Variance Inflation level			
		5%	10%	50%	100%	5%	10%	50%	100%	5%	10%	50%	100%	5%	10%	50%	100%
2	2	-160.78	-160.78	-161.50	-161.50	-150.28	-150.28	-151.00	-151.00	-156.53	-156.53	-157.24	-157.24	-160.78	-160.37	-161.08	-161.08
3	3	-165.50	-165.50	-166.74	-166.74	-152.38	-152.38	-153.62	-153.62	-160.19	-160.19	-161.43	-161.43	-165.50	-164.88	-166.12	-166.12
4	4	-161.87	-161.23	-159.38	-164.74	-146.12	-145.48	-143.63	-149.00	-155.49	-154.86	-153.01	-158.37	-161.87	-160.35	-158.50	-163.86
5	5	-161.23	-161.23	-159.95	-155.17	-142.86	-142.86	-141.58	-136.80	-153.79	-153.79	-152.51	-147.73	-161.23	-160.04	-153.98	-153.98
6	6	-162.63	-162.63	-160.06	-155.05	-141.63	-141.63	-139.06	-134.05	-154.12	-154.12	-151.56	-146.55	-162.63	-161.08	-158.76	-153.50
7	7	-161.58	-161.58	-161.58	-157.39	-137.96	-137.96	-137.96	-133.77	-152.02	-152.02	-152.02	-147.83	-161.58	-159.63	-159.63	-155.44
8	8	-162.42	-162.42	-162.42	-160.27	-136.17	-136.17	-136.17	-134.02	-151.79	-151.79	-151.79	-149.64	-162.42	-160.00	-160.00	-157.86
9	9	-162.78	-162.78	-162.78	-161.16	-133.90	-133.90	-133.90	-132.28	-151.09	-151.09	-151.09	-149.46	-162.78	-159.85	-159.85	-158.22
10	10	-162.89	-162.89	-162.89	-160.72	-131.39	-131.39	-131.39	-129.22	-150.14	-150.14	-150.14	-147.97	-162.89	-159.39	-159.39	-157.22
11	11	-157.28	-157.28	-157.28	-154.35	-123.15	-123.15	-123.15	-120.23	-143.46	-143.46	-143.46	-140.53	-157.28	-153.14	-153.14	-150.21
12	10	-160.88	-160.88	-160.88	-158.54	-129.39	-129.39	-129.39	-127.04	-148.13	-148.13	-148.13	-145.78	-160.88	-157.38	-157.38	-155.03

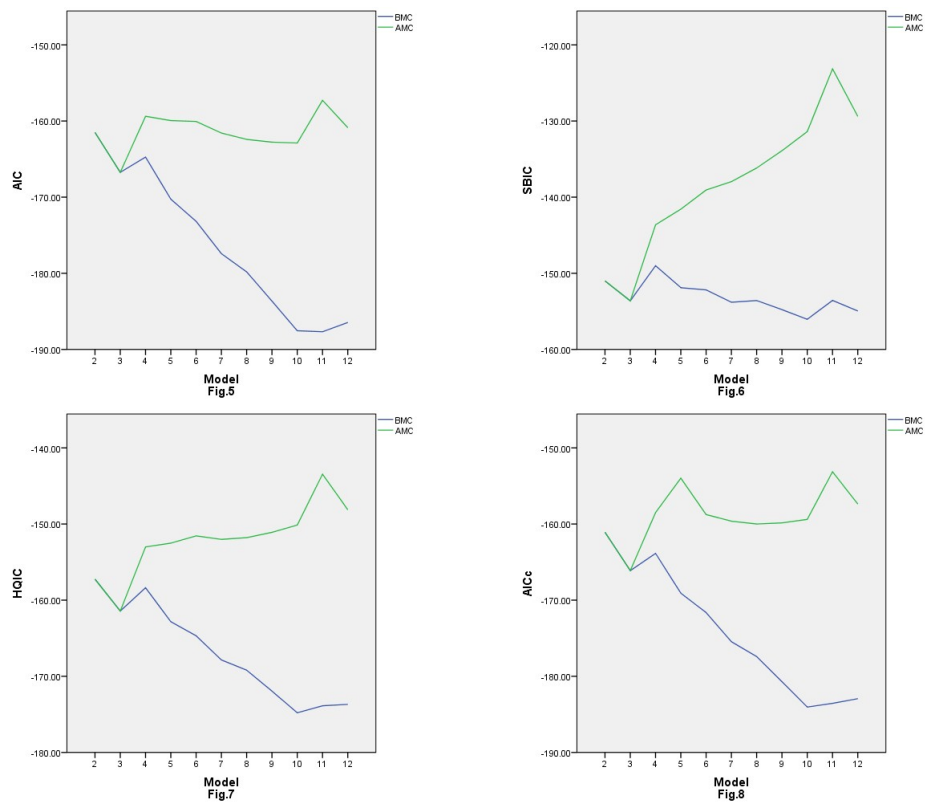


Figure 2: Line plot shows the values of Modified Versions of Information Criteria for different Models at 50% inflation in error variance

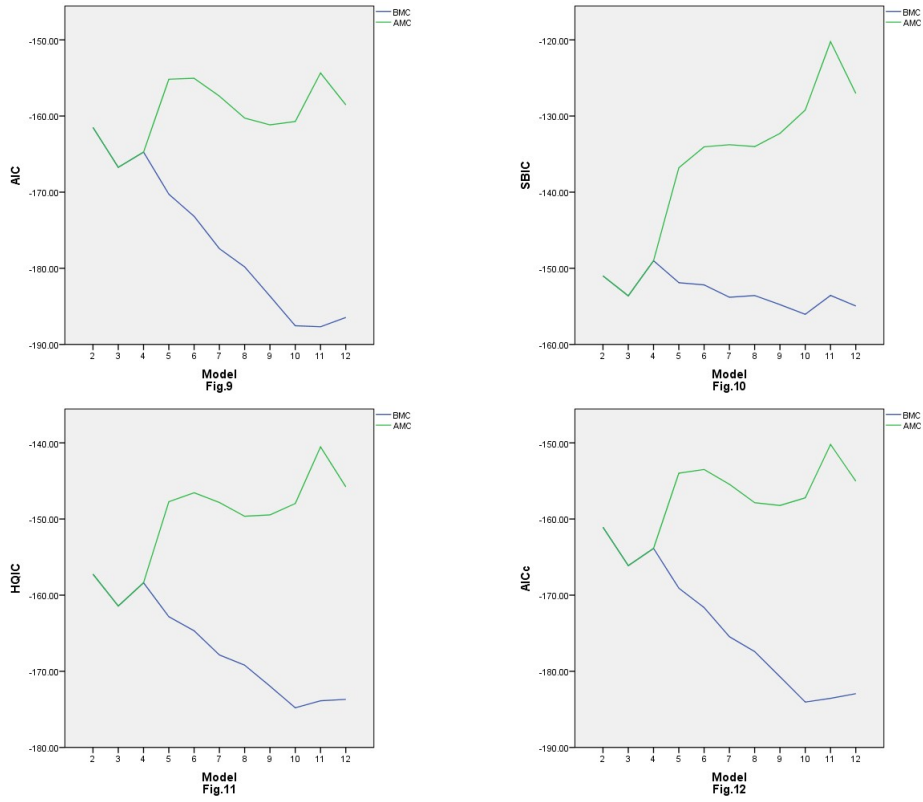


Figure 3: Line plot shows the values of Modified Versions of Information Criteria for different Models at 100% inflation in error variance

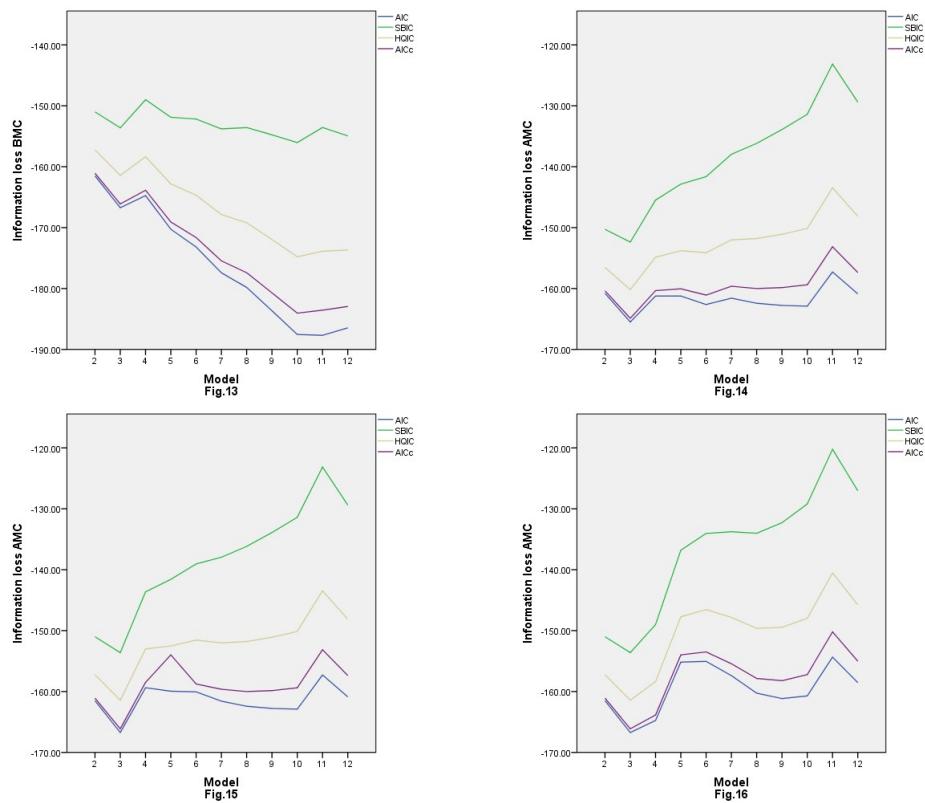


Figure 4: Line plot shows the Information Loss for different Models BMC and AMC at 5% and 10%, 50%, 100% inflation in error variance

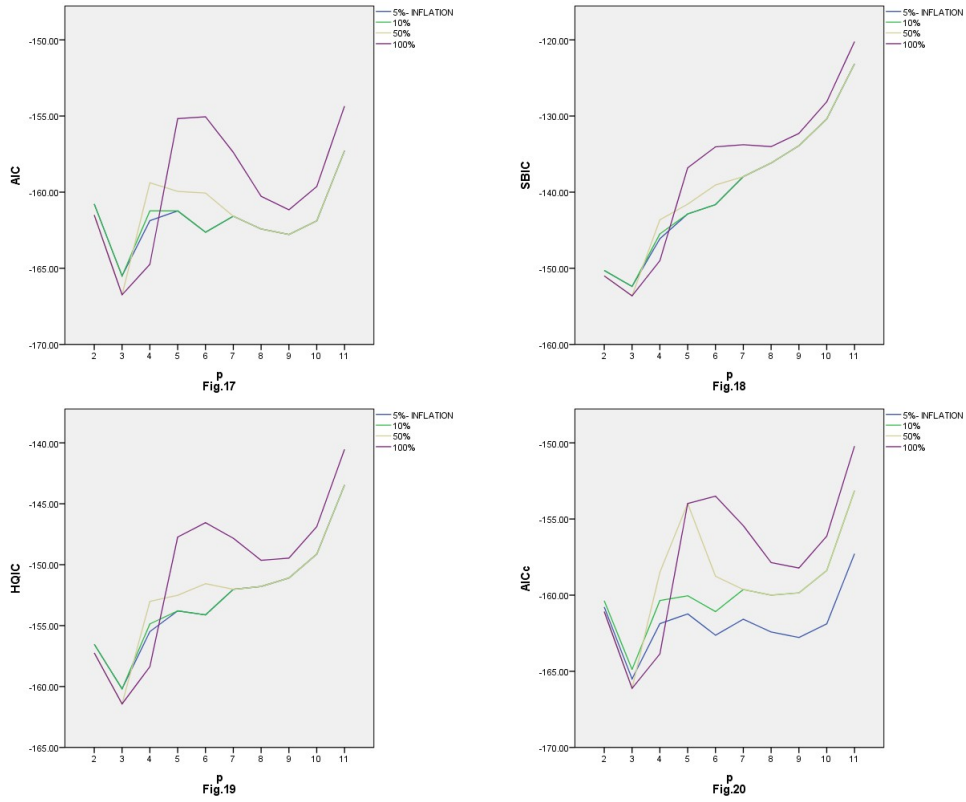


Figure 5: Line plot shows the Information Loss at 5%, 10%, 50%, and 100% inflation in error variance AMC based on P (no.of independent variables)

From the previous analysis, the authors first checked the significance of multicollinearity effect for the 12 fitted stepwise regression models. We left the results of the model 1 because it only involves one independent variable and there is no need to test the significance of multicollinearity effect. The result presented from Table-2 to 7 includes the tolerance of the independent variable, sample variance inflation factor (vif), estimated inflated error variance ($\widehat{\sigma_{INF(e)}^2}$) of the fitted model by using j th independent variable and the result of F -test for multicollinearity identification. For the model 2, the authors identified, that the result of F -test shows that the error variance of this model is inflated up to 10% due to the multicollinearity effect between the independent variables X38 and X2. Similarly, almost all the fitted models experienced a multicollinearity problem with the inflation in error variance upto 10%. As far as model 4 is concern, the F -test reveals that the estimated error variance was inflated at 50% level due to the multicollinearity effected by the independent variable X19. Similarly, from model 5 to 12, based on the result of F -test, the authors identified that the error variance of the fitted model was inflated upto 50% due to the problem of multicollinearity created by the independent variables at 1% and 5% significance level respectively. Moreover, model 5 to 12, affected severely due to multicollinearity effect among independent variables with a 100% inflation level in the error variances. From the above discussion, the authors found, due to the inflation of the error variances in the fitted models, the estimated variances of the regression co-efficients will definitely increase and it also threaten the stability of the least squares estimates of regression co-efficients. In the same manner, the authors also found another interesting result extracted from the previous analysis. The impact of multicollinearity distorts the model selection decision and from table-8 to 11 with graphs exhibits the transitions in the model selection decision at 5%, 10%, 50% and 100% inflation level in error variances based on 4 frequently used information criteria such as AIC, SBIC, HQIC and AICc. As far as AIC is concern, it suggests model 11 is the best before carrying multicollinearity correction because the information loss between the true and the fitted model is minimum when compared to remaining competing models. As far as SBIC, HQIC and AICc are concern, model 10 is the best before the incorporation of multicollinearity correction. But after the incorporation of the multicollinearity effect in model selection criteria, the selection decision varied. From table-8, 9, 10 and 11, if the error variance was inflated upto 5%, 10%, 50% and 100% respectively, the result of the modified information criteria such as AIC, SBIC, HQIC and AICc recommends, model-3 is the best in minimum information loss when compare to remaining competing models. Hence the simulation results proves that there is a definite transition in model selection after removing the effect of multicollinearity from the fitted models.

6 Conclusion

From the previous sections, the authors explained how the multicollinearity influences the fitness of the model and how it impacts the model selection decision. For this the authors, utilized F -test for scrutinizing the identification of multicollinearity based on the estimated inflated and uninflated error variance of a fitted regression model. While

taking a model selection decision, the statisticians should consider the problems in the fitted models. Multicollinearity is not violating the assumption of least squares or maximum likelihood approach but some care should be given while selecting a model affected by multicollinearity. The authors tactically penalized a model with multicollinearity and they showed the inflation in various levels of the error variances among the competing models. On the other hand, while taking a model selection decision based on information criteria, these are incapable of reflecting the problems in the fitted models. For this, the authors made a multicollinearity correction in GIC and incorporate the multicollinearity effect by replacing the inflated error variance instead of using the un-inflated error variance of a model. Hence the authors emphasize while selecting a model, statisticians should consider the problems in the fitted model and we should incorporate the consequences of this problem in the model selection yardsticks. As far as least square problems are concern, the authors suggested, statisticians to incorporate the problems such as Heteroskedasticity, autocorrelation, outliers and influential points in the model selection decision, only then we can do an unbiased model selection process. In future, more rigorous simulation experiments may conduct by fixing the variance inflation level between 1% to 100% with an interval of 0.1% helps the statisticians to find an exact variance inflation level of the error variance. Moreover, based on the relationship between the F-ratio and sample variance inflation factor (vif), we can derive the exact sampling distribution of the vif and its properties will motivates the statisticians to take the multicollinearity problem to the next level.

References

- Bowerman, B. L., O'Connell, R. T. and Richard, T. (1993), *Forecasting and Time Series: An Applied Approach* Belmont, CA Wadsworth.
- Draper, N. and Smith, H. (1981), *Applied Regression Analysis* New York: Wiley.
- Maddala, G. S. (1977), *Econometrics* New York: McGRAW-Hill Book Company.
- Neter, J., Wasserman, W. and Kutner, M. H. (1989), *Applied Linear Regression Models* Homewood, IL Richard D. Irwin.
- B. Efron,(2004), The estimation of prediction error: covariance penalties and cross-validation *J. Amer. Statist. Assoc.*, **99**,619–632
- Srivastava,M. S. (2002), *Methods of Multivariate Statistics* John Wiley and Sons,NewYork.